Using machine learning methods to predict dry matter intake from milk mid-infrared spectroscopy data on Swedish dairy cattle

Suraya Mohamad Salleh, Rebecca Danielsson and Cecilia Kronqvist

Supplementary Material

Text S1

Materials and methods

Data collection

Total raw feed intake was recorded with BioControl's System for Controlling and Recording Feed Intake V 2.0 (CRFI) for forages and with concentrate dispensers (FSC400, DeLaval International AB). Silage samples were taken from the silo for dry matter (DM) analysis five days per week and combined into one sample per fortnight for further analysis. The DM of forages was analysed following the Nordic Feed Evaluation System (NorFor) (Volden, 2011). For concentrate, a DM concentration of 880 g/kg was assumed.

Milk samples from morning and evening milking were taken fortnightly, preserved with Bronopol, refrigerated and analysed within three days using MIR spectroscopy (CombiScope FTIR 300 HP, Delta Instruments B. V., Drachten, the Netherlands). The spectrum for one morning sample for each individual and sampling bout was used in this study. The data for each full MIR spectrum consisted of absorbance in 935 wavenumbers ranging from 397.307 to 4000.071 cm^{-1}. The daily milk yield readings were averaged over the seven days prior to milk sampling.

PLS

In PLS regression, the pls v2.8-0 package (Wehrens & Mevik, 2007) was used to develop the models predicting DMI. The pls package was developed for multivariate data like the MIR spectra data with 935 variables, which were assigned as a matrix to facilitate the model functions.

Prior to validation, the model was initially tested with a maximum of 15 components or latent variables (LVs) and with the leave-one-out cross-validation (LOOCV) method. The number of LVs was selected based on the lowest root mean squared error of calibration (RMSEC), as the lowest number of LVs where RMSEC did not decrease on adding another LV. The details of the results can also be found in this Supplementary Table S2.

SVM

In SVM regression, the e1071 package version 1.7-9 (Karatzoglou *et al.*, 2006) was used. In SVM regression model development, the *cost* function was set at 5 based on validation after fine tuning values between 0.01 and 10 in the training set with 10-fold cross validation, and *gamma* function was set at default value (0.001) in the training dataset. The radial basis function (RBF) kernel was chosen, as it is suitable for modelling non-linear relationships between variables.

RF

In RF regression, the randomForest package version 4.6-14 (Breiman, 2001) was used and the main hyperparameters, which were ntree (number of trees) set at 500 and mtry (the size of variable subset), were chosen automatically when performing the RF regression model function.

Text S2

Results

Descriptive statistics on the data used in the analysis are presented in Table S1 and Figure S1. Table S1 shows the range, mean and standard deviation of total DMI, daily MY, total forage DMI and total concentrate DMI for each cow in a day (kg/day) for the training and test dataset. Figure 1(a) shows average daily MY (kg/day), while Figure S1(b) shows average DMI (kg/day) and concentrate DMI (kg/day) per cow across 180 days in milk. Milk production increased from 3-50 DIM and started to decrease gradually from approximately day 50 to the end of available data (DIM 180). The forage proportion in the diet ranged between 24 and 90 %. In total, 364 dairy cows were included, representing a total of 449 lactations.

Table S1: Descriptive statistics on dry matter intake (DMI) data for all cows involved in the model training and test datasets

	Training data (2017-2020); n= 1323				Test data (2021); n=471			
Parameter	Min	Max	Mean	SD	Min	Max	Mean	SD
Total DMI, kg/day	9.4	37.1	25.0	4.3	8.6	33.5	22.7	4.0
Daily MY, kg/day	4.4	58.2	35.5	8.4	4.8	68.2	40.0	10.6
Forage DMI, kg/day	2.7	32.0	14.9	4.6	4.1	26.4	13.2	3.6
Concentrate DMI, kg/day	1.7	16.8	10.1	3.5	1.6	16.8	9.6	4.7

Min: Minimum; Max: Maximum; SD: Standard deviation; MY: Milk yield.

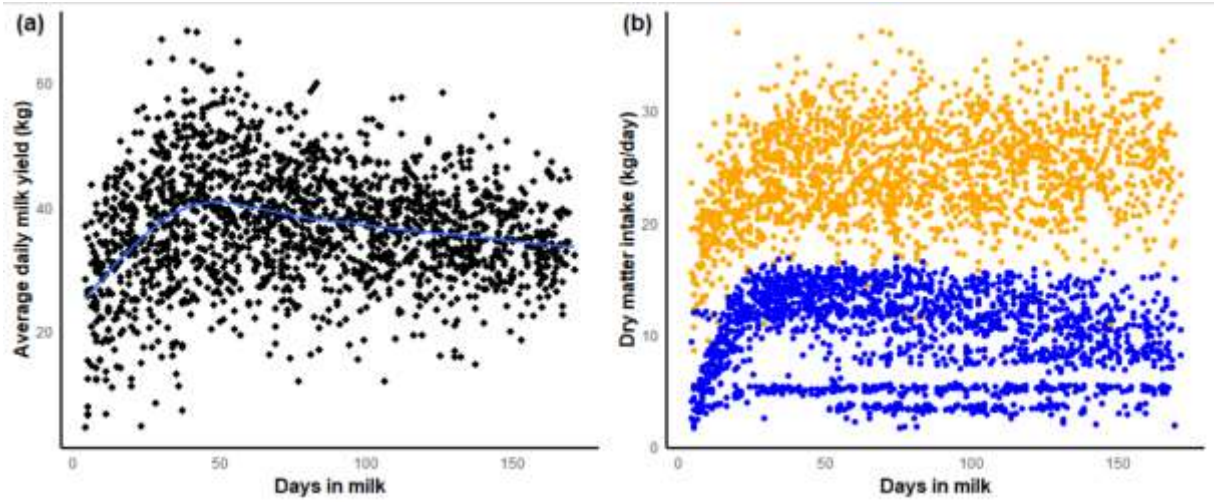

Figure S1: (a) Scatter plot (black points) of average daily milk yield values (kg/day), with the overall average shown as a blue line, and (b) scatter plot of total dry matter intake (orange points) and concentrate dry matter intake (blue points) in the first 180 days in milk.

Although the coefficient of determination (R^2=0.52) of the test dataset was moderate-high in the SVM regression approach, the range of predicted DMI values was wider using PLS regression (10-34 kg/day) compared with the other two approaches (15-32 kg/day and 16-29 kg/day for SVM and RF, respectively), Supplementary Figure S2.

Table S2: Prediction accuracy of PLS regression analysis. Optimum number of latent variables, coefficient of determination (R^2) for calibration model and validation dataset, RMSEP and MAE between predicted and actual observations of DMI in kg/day

PLS	LVs	R^2 (CV)	R^2 (test)	RMSEP	MAE
DIM 3-180					
935 MIR	10	0.23	0.19	3.67	2.96
MY	-	0.35	0.31	5.03	4.19
Conc	-	0.10	0.46	3.73	3.00
935 MIR + MY	13	0.51	0.43	3.19	2.48
935 MIR + MY+ Lact stage	15	0.50	0.44	3.07	2.39
935 MIR + MY+ Lact stage+Par	15	0.50	0.44	3.09	2.42
935 MIR + Conc	10	0.40	0.44	3.86	2.99
935 MIR + MY + Conc	10	0.49	0.62	2.71	2.13
DIM 3-30					
MIR	10	0.34	0.40	3.19	2.65
MY	-	0.41	0.44	3.44	2.63
Conc	-	0.31	0.58	2.92	2.29
MIR + MY	15	0.56	0.55	2.96	2.30
935 MIR + Conc	10	0.50	0.58	2.88	2.21
935 MIR + MY + Conc	10	0.62	0.65	2.65	2.14
DIM 30 - 180					
MIR	12	0.18	0.20	3.49	2.87
MY	-	0.28	0.24	5.18	4.31
Conc	-	0.05	0.49	3.90	3.21
MIR + MY	14	0.39	0.40	3.10	2.42
935 MIR + Conc	10	0.34	0.43	3.99	3.13
935 MIR + MY + Conc	10	0.40	0.60	2.62	2.05

PLS: Partial least square regression; MIR: milk mid-infrared spectroscopy; LVs: Latent variables; MY: average daily milk yield; DIM: days in milk; Conc: concentrate DMI; Lact stage: lactation stage; Par: parity; CV: cross validation; RMSEP: root mean square error of prediction; MAE: mean absolute error

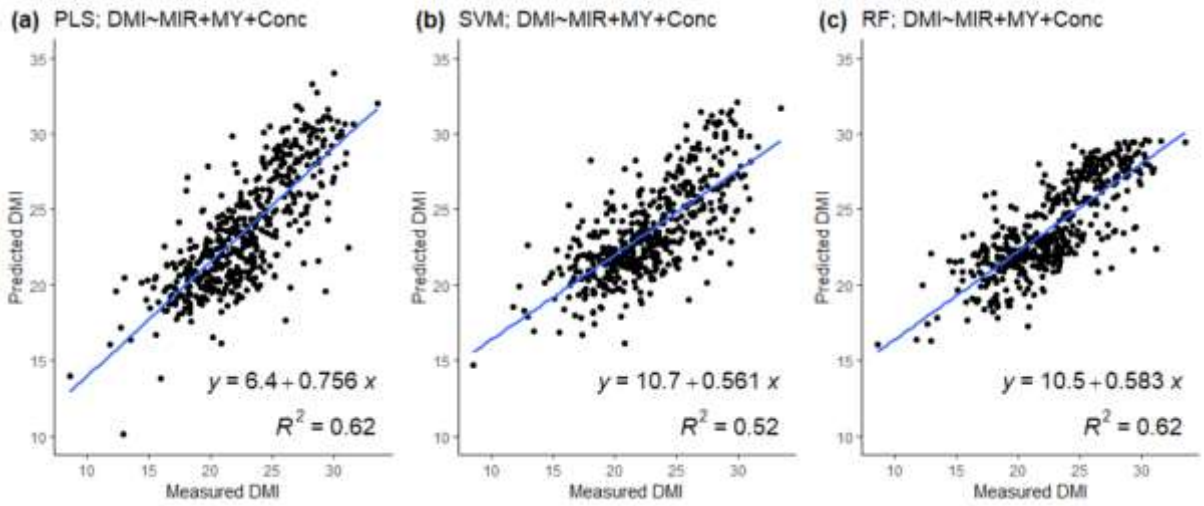

Figure S2: Plots of predicted dry matter intake (DMI, g/day) against measured DMI (kg/day) of the validation/test set for models with milk mid-infrared (MIR) + milk yield (MY) + concentrate (Conc) as the predictor (days in milk (DIM): 3-180). Plots show the regression line and coefficient of determination (R^2) value for (a) partial least squares (PLS) regression, (b) support vector machine (SVM) regression and (c) random forest (RF) regression